# 503project-EDA

Maha Jayapal

6/8/2022

Loading the packages

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.3

## Loading required package: lattice
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.3
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.1.3

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:e1071':
##
##     impute

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(corrplot)
```

```
## Warning: package ’corrplot’ was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
library(plyr)
```

```
## Warning: package ’plyr’ was built under R version 4.1.3
```

```
##
## Attaching package: ’plyr’
```

```
## The following objects are masked from ’package:Hmisc’:
##
##     is.discrete, summarize
```

```
library(pROC)
```

```
## Warning: package ’pROC’ was built under R version 4.1.3
```

```
## Type ’citation("pROC")’ for a citation.
```

```
##
## Attaching package: ’pROC’
```

```
## The following objects are masked from ’package:stats’:
##
##     cov, smooth, var
```

```
stroke<- read.csv('c:\\maha\\503\\healthcare-dataset-stroke-data.csv', header = TRUE)
```

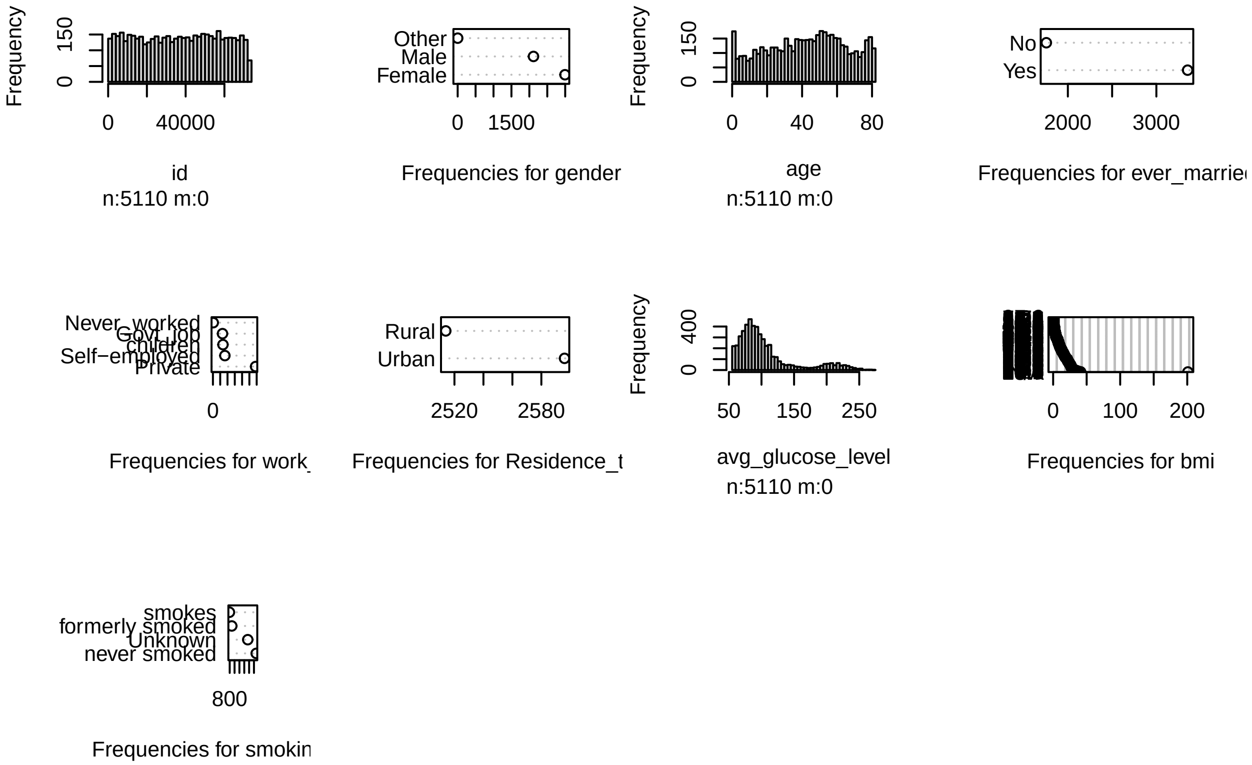
```
str(stroke)
```

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
dim(stroke)
```

```
## [1] 5110   12
```

```
hist.data.frame(stroke)
```



```
skewness(stroke$avg_glucose_level)
```

```
## [1] 1.571361
```

```
skewness(stroke$age)
```

```
## [1] -0.1369789
```

```
table(stroke$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
table(stroke$hypertension)
```

```
##
##    0    1
## 4612  498
```

```
table(stroke$heart_disease)
```

```
##
##    0    1
## 4834  276
```

```
table(stroke$ever_married)
```

```
##
##   No  Yes
## 1757 3353
```

```
table(stroke$work_type)
```

```
##
##      children      Govt_job  Never_worked       Private Self-employed
##           687           657            22          2925           819
```

```
table(stroke$Residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```
table(stroke$smoking_status)
```

```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

```
table(stroke$stroke)
```

```
##
##    0    1
## 4861  249
```

```
stroke$gender <- as.numeric(as.factor(stroke$gender))
stroke$ever_married <- as.numeric(as.factor(stroke$ever_married))
stroke$work_type <- as.numeric(as.factor(stroke$work_type))
stroke$Residence_type <- as.numeric(as.factor(stroke$Residence_type))
stroke$smoking_status <- as.numeric(as.factor(stroke$smoking_status))
stroke$bmi <- as.numeric(stroke$bmi)
```

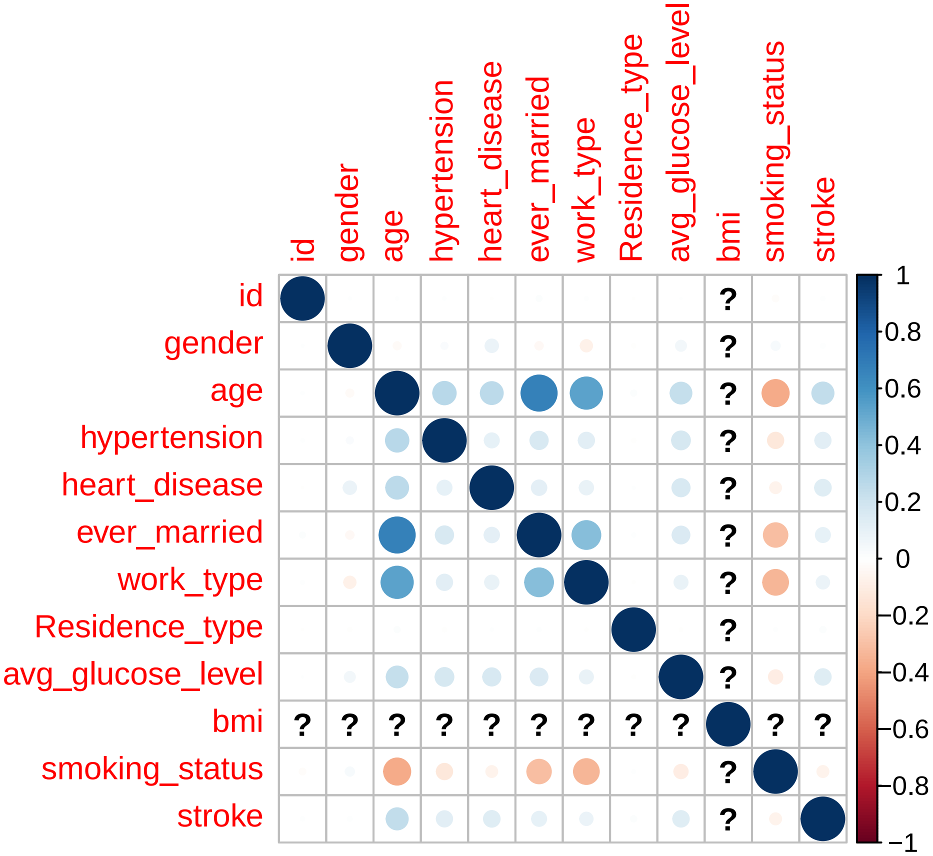
```
## Warning: NAs introduced by coercion
```

```
#stroke$stroke <- as.factor(stroke$stroke)
```

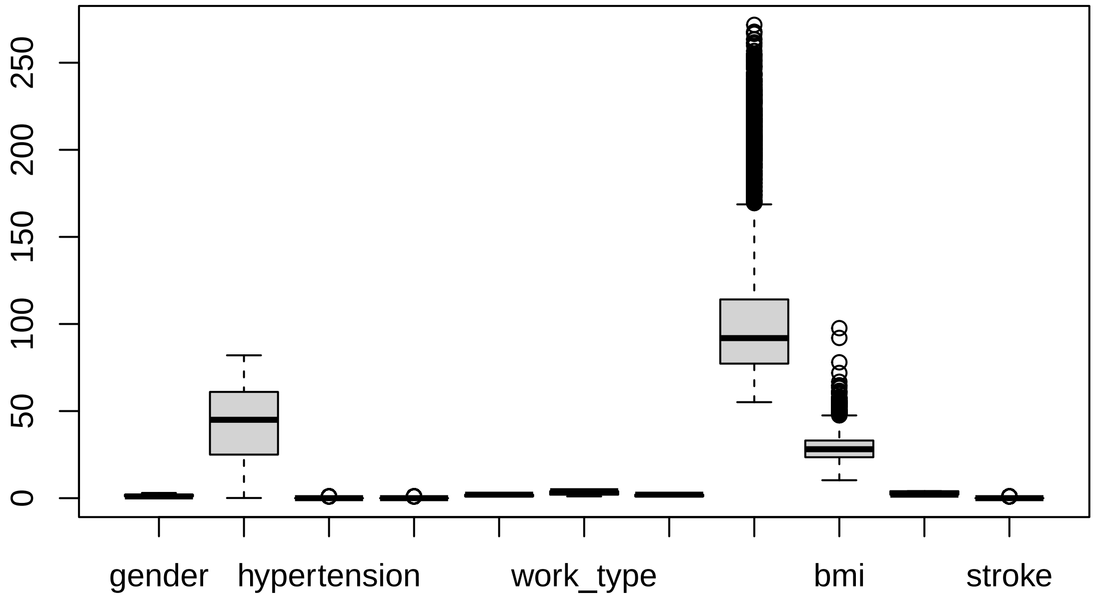
```
str(stroke)
```

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : num  2 1 2 1 1 2 2 1 1 1 ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : num  2 2 2 2 2 2 2 1 2 2 ...
##  $ work_type        : num  4 5 4 4 5 4 4 4 4 4 ...
##  $ Residence_type   : num  2 1 1 2 1 2 1 2 1 2 ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : num  1 2 2 3 2 1 2 2 4 4 ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
corrplot(cor(stroke))
```



```
boxplot(stroke[,-1])
```


0
50
100
150
200
250
gender
hypertension
work_type
bmi
stroke